Cognitive and Corpus-Based Approaches to Language Processing

Assessing Reading Development through Finger-Voice Span: a Comparative Study of Bulgarian and Italian

Andrea Nadalini[1], Claudia Marzi[1], Marcello Ferro[1], Vito Pirrelli[1],
Maria Todorova[2], Valentina Stefanova[2], Tsvetana Dimitrova[2],
Hristina Kukova[2] and Svetla Koeva[2]

[1]Institute for Computational Linguistics "A. Zampolli", Italian National Research Council
{andrea.nadalini,claudia.marzi,marcello.ferro,vito.pirrelli}@cnr.it

[2]Institute for Bulgarian Language, Bulgarian Academy of Science
{maria,valentina,cvetana,hristina,svetla}@dcl.bas.bg

Abstract

This study investigates finger–voice span (FVS) as an index of reading development in primary school children, through a cross-linguistic comparison between Italian and Bulgarian. FVS measures the spatial distance, in characters, between the finger position in a finger-point oral reading session and the voice onset of the word being vocalized during the same session, providing insight into the coordination of visual, motor, and linguistic processes during reading. Data were collected from 231 children in grades 2-5, during oral reading tasks administered through the ReadLet application. Finger movements and voice articulation were synchronously recorded for subsequent alignment and analysis. To ensure reliable temporal alignment for Bulgarian, a subset of recordings was manually annotated at the token level to evaluate an Automatic Speech Recognition (ASR) pipeline adapted from the Italian version. The ASR system, based on WhisperX and enhanced with energy-based segmentation and dual alignment, provided reliable temporal alignment, supporting its use for large-scale analyses. FVS measures were extracted from more than 44,000 tokens and analysed using linear mixed models. Results revealed that both grade and language significantly influenced FVS. Italian children exhibited higher FVS values than their Bulgarian peers across all grade levels, while FVS increased with grade in both languages, reflecting shared developmental changes in reading coordination. These results provide evidence of shared developmental changes in FVS across languages, alongside language-specific differences. The study highlights the potential of the finger-tracking technique combined with ASR for a scalable, cross-linguistic assessment of reading proficiency in naturalistic settings.

Keywords: reading, finger-tracking, reading, finger-voice span (FVS), automatic speech recognition (ASR), Bulgarian, Italian.

1 Introduction

Oral reading of connected text relies on the tight coordination between ocular and articulatory movements. Eye movements are faster and more flexible than articulatory movements. They typically scan a text forward and backward using a variety of strategies, such as forward saccades, regressions, refixations, and word skipping. In contrast, the articulatory output of a skilled reader is characterised by smooth, continuous articulation, with few repetitions or disfluencies and only occasional pauses at linguistically complex boundaries (Inhoff et al., 2011; De Luca et al., 2013; Ziaka and Protopapas, 2025).

To coordinate these processes effectively, readers rely on a fine-tuned motor control strategy. This involves temporarily storing decoded units in a phonological buffer, integrating them into larger prosodic structures, and planning and executing articulation based on the buffered material (Laubrock and Kliegl, 2015).

This dynamic can be investigated by measuring the eye-voice span (EVS), defined as the temporal or spatial distance between the onset of the first fixation on a word and its articulation. Skilled readers can maintain more information in the phonological buffer and adjust their reading pace to larger linguistic units, such as chunks and phrases, reflecting efficient prediction, fluent decoding, and effective speech planning. Less proficient readers, by comparison, tend to show a smaller and less variable EVS, with eye movements more tightly aligned to oral output, reflecting slower word-by-word processing. By capturing the coordination between eye movements, lexical access, and speech planning, EVS provides an informative index of reading proficiency (Nadalini et al., 2023; Marzi et al., 2025).

It has recently been shown (Taxitari et al., 2021; Ferro et al., 2024) that finger-tracking data provide a compelling alternative to eye-tracking evidence, enabling high–temporal resolution measurement of reading behaviour in naturalistic settings. Finger-tracking captures the timing and trajectory of the reader's index finger as it slides along the lines of a written text during finger-point reading, yielding fine-grained evidence of processing without requiring specialised laboratory equipment. In this study, we used the ReadLet infrastructure, a cloud-based platform delivering reading tasks via a tablet application, which enables simultaneous recording of finger movements and voice articulation in oral reading (Ferro et al., 2018).

Crucially, finger-tracking measures have been shown to strongly correlate with established eye-tracking indices, such as fixation durations and saccadic patterns, in both adult and developing readers (Crepaldi et al., 2022; Marzi et al., 2020). In addition, finger-voice span (FVS) captures a coordination pattern similar to the eye–voice span. In early readers, haptic movements are tightly coupled with gaze, indicating that finger pointing supports visual attention and sequential scanning of letters and syllables (Mesmer and Lake, 2010). As reading proficiency develops, however, the role of the finger shifts from supporting visual alignment to tracking articulation timing, with its movements becoming more synchronised with the speech stream (Nadalini et al., 2023; Marzi et al., 2025).

This convergence suggests that finger movements can reliably reflect underlying oculomotor control and lexical processing. In addition, the portability, scalability, ease of implementation, and relatively low cost of finger-tracking make it particularly well suited for large-scale studies in educational contexts and across diverse languages, such as Bulgarian and Italian (Lento et al., 2024; Marzi et al., 2026). These features support broader access to developmental reading data and facilitate cross-linguistic comparisons, with potential applications in educational practice and literacy policy (Nadalini et al., 2025).

Here, we present a comparative analysis of FVS in Italian and Bulgarian primary school pupils. To ensure reliable token-level alignment in Bulgarian data, we first evaluated an automatic speech recognition (ASR) system by comparing its output against manually time-stamped recordings, using the Italian-developed system as a reference (Srivastava et al., 2026). Thus, this study addresses two main research questions: (1) Does Finger-Voice Span increase with reading development in Bulgarian? (2) Does the developmental trajectory of FVS differ between Bulgarian and Italian?

2 Participants

The dataset consists of 231 primary school pupils from grades 2-5, recruited in Pisa (Italy, 109 recording sessions) and Sofia (Bulgaria, 122 recording sessions). Participation required parental consent and followed no selection criteria based on reading proficiency. All participants had normal or corrected-to-normal vision, no reported hearing impairments, and no diagnosed neurodevelopmental disorders, learning disabilities, or cognitive impairments. Participants were native or dominant speakers of the language tested at their respective location. Demographic characteristics of both Bulgarian and Italian participants are reported in Table 1.

Grade	N	Female	Male	Mean age (sd)
		BG		
2	20	7	13	8.64 (0.33)
3	38	16	22	9.57 (0.37)
4	32	16	16	10.50 (0.35)
5	32	24	8	11.50 (0.37)
		IT		
2	18	9	9	7.65 (0.35)
3	22	14	8	8.78 (0.20)
4	32	15	17	9.74 (0.33)
5	37	17	20	10.00 (2.50)

Table 1: Distribution of Bulgarian (BG) and Italian (IT) participants by grade level. Age is reported in decimal months.

3 Material and Methods

Data were collected using a 10.1-inch Android tablet. Speech were recorded with a bidirectional electret microphone with a 150-6800Hz frequency range and a sensitivity of -47dBV/Pa. The audio signal was further resampled at 48000Hz, 16bit, stereo and compressed to audio-WEBM format at 128kbps.

The reading materials consisted of narrative texts designed to be comparable across languages in terms of length and lexical, morphological, and syntactic properties. The same set of narratives was originally written in Italian and subsequently translated into Bulgarian following procedures aimed at preserving structural and linguistic equivalence (Pirrelli and Koeva, 2024). Each narrative was structured into five episodes with gradually increasing length and complexity. Children read a grade-specific subset of each narrative, ranging from two episodes in grade 2 to the full five-episode set in grade 5. All texts were processed with language-specific annotation tools (Dell'Orletta et al. 2011, for Italian; Karagiozov et al. 2011, for Bulgarian). Lexical frequency measures were derived from Subtlex (Crepaldi et al., 2016) and the Bulgarian National Corpus (Koeva et al., 2012). Summary statistics for the reading materials are reported in Table 2.

	Ep. 1	Ep. 2	Ep. 3	Ep. 4	Ep. 5
			BG		
Word Count	119.75 (7.14)	133.25 (14.13)	140.75 (6.85)	151.00 (12.99)	155.50 (9.33)
Word Frequency	5.42 (1.32)	5.27 (1.46)	5.18 (1.51)	5.31 (1.47)	5.31 (1.47)
Word Length	4.57 (2.28)	4.78 (2.68)	4.88 (2.82)	4.80 (2.80)	4.82 (2.94)
Sentence Count	13.75 (0.50)	11.50 (0.58)	10.00 (0.00)	8.00 (0.82)	7.50 (0.58)
Sentence Length	8.70 (0.27)	11.59 (1.15)	14.08 (0.68)	18.92 (1.09)	20.82 (1.99)
Dependency Length	1.73 (1.42)	2.03 (2.09)	2.22 (2.22)	2.30 (2.45)	2.59 (2.94)
			IT		
Word Count	141.00 (0.82)	152.00 (0.82)	166.00 (4.83)	169.75 (3.20)	178.00 (7.96)
Word Frequency	5.76 (1.27)	5.45 (1.47)	5.46 (1.49)	5.41 (1.51)	5.38 (1.53)
Word Length	4.21 (2.17)	4.66 (2.46)	4.71 (2.53)	4.87 (2.76)	5.07 (3.01)
Sentence Count	14.00 (0.00)	11.50 (0.57)	10.00 (0.00)	8.00 (0.82)	7.50 (0.58)
Sentence Length	10.07 (0.00)	13.33 (0.67)	17.00 (0.00)	21.80 (2.25)	23.57 (1.81)
Dependency Length	1.57 (1.39)	1.78 (1.79)	1.93 (2.04)	2.06 (2.43)	2.33 (2.97)

Table 2: Descriptive statistics (mean and standard deviation) for Bulgarian (BG) and Italian (IT) reading texts by text episodes.

4 Manual Annotation of Bulgarian data

Manual annotation was carried out on a subset of Bulgarian audio recordings, corresponding to 40 text episodes out of a total of 387 recordings. The sample included ten recordings for each grade level (grades 2-5). Annotators were asked to mark the time onset and offset of each vocalized word in an audio recording, and type the transcript into the resulting time interval using Praat's TextGrid editor (Boersma and Weenink, 2024).

The subset of reading sessions was selected to ensure coverage of different text episodes and a balanced distribution across male and female participants. Each audio file was time-aligned with a tokenized version of its corresponding written text, with tokens being defined as individual orthographic words. Annotation included the creation of synchronised tiers for audio–text alignment, ensuring correspondence between the acoustic signal and the text. Alignment was supported by spectrogram and waveform visualisations, with appropriately set frequency ranges and temporal resolution. This enabled precise localisation of acoustic events and improved discrimination of temporally adjacent segments. Token boundaries were manually marked based on acoustic criteria, taking into account phoneme transitions, coarticulation effects, and reductions typical of spontaneous speech. In cases of ambiguous boundaries, a conservative strategy was adopted, with transitional segments assigned to the relevant token. This typically occurred in fast reading conditions, when prepositions or conjunctions merge with adjacent words, making exact boundary identification difficult.

Additional labels with annotators' comments were applied to mark occurrences of word repetitions, fragmented or interrupted words, hesitations, filled pauses, and instances of syllabification produced during reading. These phenomena can be further categorised according to established typologies of speech disfluencies (Zhang, 2020), including:

- Repetitions – immediate reiteration of a word, words, parts of words during reading.
- Self-interruptions / broken words – abrupt cessation of a word before completion.
- Restarts – abandonment of an initiated word or phrase followed by a new attempt at its pronunciation.
- Filled pauses – vocalized hesitation markers such as uh, um, or other non-lexical fillers.
- Unfilled pauses – silent intervals exceeding a predefined duration threshold.
- Prolongations – atypically extended articulation of a sound.
- Syllabification – deliberate or unintentional production of a word in separate syllables.
- Incorrect pronunciation – deviations from the expected phonological realisation of a word.

Description of the annotation principles, annotation labels and representative examples is provided in Appendix A.

5 The ASR algorithm

We used the ASR pipeline originally developed by Srivastava et al. (2026), built on the WhisperX framework (Bain et al., 2023). The first step of the pipeline is an energy-based pause detection algorithm that segments an input waveform into prosodically coherent units using energy (loudness) thresholds. Prosodic units are then transcribed using a pre-trained, language-specific Whisper Large V3 model, which produces segment-level transcriptions. These transcriptions are subsequently normalized for punctuation and diacritics before being passed to a dual alignment module. The alignment module features separate optimization components for token onset (Grosman, 2022) and offset (Conneau et al., 2020) time alignments. The final step of the algorithm is a token-level post-processing stage that merges these alignments, correcting systematic temporal shifts and ensuring consistency across neighbouring tokens. The final output consists of token boundaries with precise temporal alignment augmented with acoustic and prosodic cues, explicitly accounting for disfluencies in spontaneous speech.

6 ASR evaluation

ASR performance was evaluated using multiple complementary metrics. First, we measured temporal deviations (deltas) between automatically detected boundaries and manual annotations, separately for onset and offset timestamps. Next, we computed Intersection-over-Union (IoU[1]) across token segments to assess overall temporal alignment quality. Finally, as an aggregate measure of performance, we computed the proportion of tokens whose onset and offset boundaries deviated by less than 50 ms from the manual reference (i.e., within a phonetic-scale tolerance). We refer to this metric as the Collared Matching Rate (CLMR). Results are shown in Table 3[2]. A consistent pattern emerges across all evaluation metrics, with higher temporal alignment accuracy for Italian than for Bulgarian. This is reflected in both boundary-based and overlap-based measures. Mean and median onset and offset deviations were lower for Italian (36–38 ms mean, 19–23 ms median) than for Bulgarian (64–67 ms mean, 36–44 ms median), indicating more precise token boundary localisation. This advantage is further supported by overlap-based evaluation, with higher IoU values for Italian (80%) compared to Bulgarian (72–78%). Similarly, the Collared Matching Rate (CLMR), measuring the proportion of words aligned within a ±50 ms tolerance, was substantially higher for Italian ($\approx$ 70%) than for Bulgarian ($\approx$ 36%). Despite this performance gap, median timing errors remained below 50 ms for both languages, suggesting that Whisper maintains a reasonable level of temporal precision in both cases. The observed differences may be attributed to a combination of factors, including language-specific phonetic and prosodic characteristics, differences in training data coverage for Whisper models, and greater acoustic variability in children's oral production in Bulgarian.

Overall, the magnitude of the language difference is not uniform, across metrics. While IoU and boundary deviation measures show moderate though consistent gaps between Italian and Bulgarian, the CLMR exhibits a substantially larger divergence, suggesting higher sensitivity to strict temporal tolerance criteria. This pattern suggests that small systematic shifts in token boundaries have limited impact on overlap-based measures but a stronger effect on tolerance-based evaluation. Thus, the three metrics provide converging evidence of cross-linguistic differences, while capturing complementary aspects of alignment accuracy. As a result, these metrics suggest that while the ASR pipeline generalises across languages, its performance is modulated by language-specific properties that affect boundary detection accuracy.

7 FVS Preprocessing

For each word in the text, spatial finger-voice span (FVS) measures the spatial distance (in characters) between the finger position within a word and the onset word articulation, namely how far the finger is when the voice starts pronouncing it. Overall, FVS values were extracted for a total of 53,938 word tokens (27,357 from Bulgarian children, 26,581 from Italian children). Tokens were excluded if tracking time and speech duration fell below 35 ms or ex-

[1]For each time-stamped token, IoU is calculated by dividing the overlapping duration (intersection) between the model estimate and the ground truth, by the total combined duration of the two intervals. IoU provides a gross estimate of the overall accuracy of the model time stamps.

[2]ASR results for Italian are taken from Srivastava et al. (2026).

Metrics	Italian	Bulgarian
Mean Onset Delta (ms)	38	64
Median Onset Delta (ms)	23	44
Mean Offset Delta (ms)	36	67
Median Offset Delta (ms)	19	36
Mean IoU (%)	80.86	71.52
Median IoU (%)	86.04	77.54
Mean CLMR (%)	70.20	36.03
Median CLMR (%)	69.74	36.64

Table 3: ASR performance comparison in the two languages.

ceeded 3.5 seconds, as well as in cases where FVS exceeded 20 characters. This filtering resulted in a final dataset of 41,926 observations included in the analyses.

8 FVS Analyses

Statistical analyses were conducted in R. Separate linear mixed models were fitted for Bulgarian and Italian using the *lme4* package (Bates et al., 2015) to examine the effects of grade level grades 2-5) on spatial onset FVS. Each model included random intercepts for word token, reader and text, accounting for repeated measures and item- and participant-level variability. Significance of the fixed effects was assessed using Type III Wald chi-square tests implemented in the *car* package (Fox et al., 2007). Model results are reported in Table 4 and model predictions are shown in Figure 1.

Grade level significantly affected onset FVS in both Bulgarian (χ^2 = 11.07, df = 3, p = .011) and Italian children (χ^2 = 9.19, df = 3, p = .027), indicating developmental changes across the primary school years. The developmental trajectories, however, differed between the two languages. In Bulgarian, predicted FVS values increased progressively across grade levels. In Italian, FVS also increased from Grade 2 to Grade 4, but showed a slight decrease in Grade 5. Although the analyses were conducted separately for the two language groups, Figure 1 descriptively shows higher predicted FVS values for Italian readers across all grade levels. The different developmental trajectories observed in Bulgarian and Italian suggest that changes in FVS may not follow a uniform developmental pattern. Rather than reflecting a simple linear increase in reading proficiency, FVS is likely shaped by multiple interacting factors, including reading experience, language-specific characteristics, and the increasing complexity of the reading material across grade levels.

BULGARIAN DATA					
Fixed Effects					
	Est.	S.E.	t val.	d.f.	p
(Intercept)	1.56	0.44	3.56	48.84	0.001
gradeLevel3	1.07	0.51	2.10	125.12	0.038
gradeLevel4	1.22	0.52	2.34	122.54	0.021
gradeLevel5	1.72	0.52	3.31	121.31	0.001
Random Effects					
Group	Parameter	Std. Dev.			
token	(Intercept)	0.82			
idSession	(Intercept)	1.75			
idDoc	(Intercept)	0.28			
Residual		3.49			
ITALIAN DATA					
Fixed Effects					
	Est.	S.E.	t val.	d.f.	p
(Intercept)	3.18	0.58	5.47	38.11	0.001
gradeLevel3	0.65	0.72	0.91	114.86	0.366
gradeLevel4	1.85	0.67	2.78	115.32	0.006
gradeLevel5	1.35	0.65	2.08	115.54	0.040
Random Effects					
Group	Parameter	Std. Dev.			
token	(Intercept)	1.27			
idSession	(Intercept)	2.14			
idDoc	(Intercept)	0.42			
Residual		3.69			

Table 4: LMM results for spatial onset Finger-Voice Span in Bulgarian and Italian separately

9 Discussion

9.1 Developmental findings

In the present study, we examined developmental changes in finger–voice span (FVS) in Bulgarian and Italian primary school children. Separate analyses revealed a significant effect of grade level in both language groups, indicating developmental changes in FVS across the primary school years. Italian children consistently exhibited higher FVS values than their Bulgarian peers across all grade levels. The developmental trajectories slightly differed between the two groups: Bulgarian children showed a steady increase in FVS across grades, whereas Italian children displayed an increase up to grade 4 followed by a slight decrease in grade 5. A supplementary model including both language groups did not show a significant language x grade interaction, suggesting that the overall effect of grade was

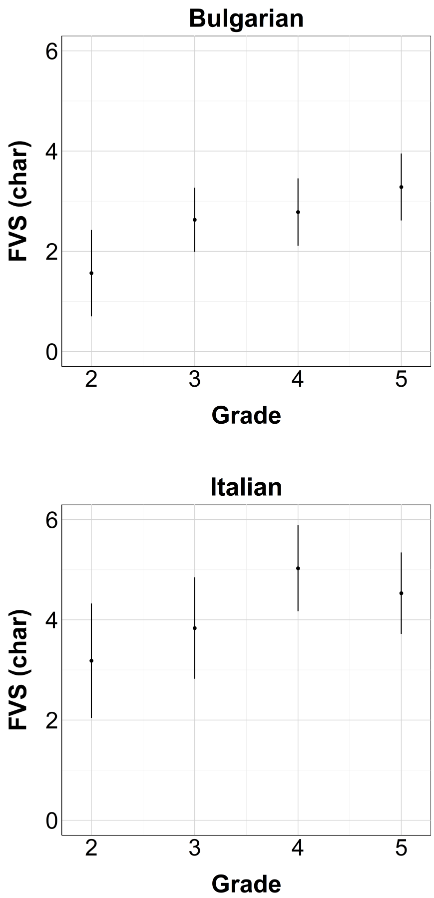

Figure 1: LMM estimates of average FVS values across grades for Bulgarian (top panel) and Italian (bottom panel) children. Error bars represent 95% prediction intervals.

broadly comparable across languages.

The tendency for Italian children to exhibit higher FVS values than Bulgarian children be due to several factors. First, it may reflect differences in lexical and morphosyntactic predictability between the Italian and Bulgarian texts. Additionally, it may be a consequence of the coarser time-stamps of Bulgarian automated transcriptions. An earlier estimated onset of word articulation (relative to the real onset) may, in fact, underestimate the distance of the finger from the voice. Finally, it may also be the outcome of different oral reading styles and teaching practices. However, the absence of a significant language x grade interaction in the combined model suggests that these cross-linguistic differences do not substantially modify the overall developmental trajectory of FVS. Despite qualitative differences in the observed patterns, both groups exhibit an overall increase in FVS with reading experience, pointing to broadly shared developmental constraints. During the early stages of reading acquisition, motor anchoring strategies such as finger pointing are thought to support visual attention and sequential decoding. As reading proficiency increases, however, this coupling gradually becomes less dominant, giving way to a more flexible coordination of visual, motor, and linguistic processes (Marzi et al., 2025).

9.2 Implications

The present findings suggest that tracking the development of finger-voice span (FVS) may provide a means of identifying both typical and atypical patterns in early reading learning during primary school years. Given that reading difficulties often emerge subtly in this period, FVS may offer a complementary measure for monitoring reading development.

From an applied perspective, the results suggest that finger-tracking can be implemented using standard tablet devices, enabling unobtrusive collection of multimodal reading data in classroom environments. Building on prior implementations in Italian schools, the present study shows that a similar protocol can be deployed in a different linguistic context with minimal adaptation. This suggests that the approach is suitable for scalable data collection and repeated assessment of reading development at both group and individual levels. Such fine-grained longitudinal data may, in fact, support learner-specific analyses that could inform personalised feedback on reading development.

9.3 Methodological considerations and limitations

An important methodological advantage of the present framework lies in the use of synchronised multimodal data streams, which allow for cross-channel constraints that can improve signal quality during processing. In particular, information from one modality can be used to mitigate noise or temporal uncertainty in another. For example, misalignment or drift in finger-tracking data can be partially corrected using temporally aligned speech signals in oral reading, jointly anchored to the text (Marzi et al., 2025). Conversely, speech data can provide coarse-grained information about the reader's position in the text. However, these benefits depend on the quality and robustness of each individual modality, and residual errors in ASR or segmentation may propagate across streams. This introduces a potential source of shared error that should be considered when interpreting fine-grained temporal measures such as FVS.

9.4 Multimodal integration and future directions

Taken together, the present findings are consistent with a model in which FVS development reflects the interplay between general maturational factors and language-related characteristics that influence baseline levels, while developmental changes across grades remain broadly comparable across languages, in line with evidence on reading speed and automatisation (Marzi et al., 2026). Future research should further disentangle the relative contribution of linguistic, educational, and age-related factors to FVS development, ideally using fully age-matched samples and more tightly controlled instructional conditions. Extending this approach to a wider range of languages would further clarify the generalisability of the observed developmental patterns.

Acknowledgments

The present study is carried out within the project National Scientific Program "Development of Scientific Research and Innovation in the System of Bulgarian Preschool and School Education" Agreement SP-13/04.11.2025

References

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

Paul Boersma and David Weenink. 2024. Praat: Doing phonetics by computer. https://www.praat.org/. Version 6.x.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition.

D Crepaldi, S Amenta, P Mandera, E Keuleers, and M Brysbaert. 2016. Frequency estimates from different registers explain different aspects of visual word recognition. In *International meeting of the Psychonomic society, Granada, Spain*, pages 5–8.

Davide Crepaldi, Marcello Ferro, Claudia Marzi, Andrea Nadalini, Vito Pirrelli, and Loukia Taxitari. 2022. *Finger Movements and Eye Movements During Adults' Silent and Oral Reading*, pages 443–471. Springer International Publishing, Cham.

Maria De Luca, Maria Pontillo, Silvia Primativo, Donatella Spinelli, and Pierluigi Zoccolotti. 2013. The eye-voice lead during oral reading in developmental dyslexia. *Frontiers in human neuroscience*, 7:696.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Marcello Ferro, Claudia Cappa, Sara Giulivi, Claudia Marzi, Oaufae Nahli, Franco Alberto Cardillo, and Vito Pirrelli. 2018. Readlet: Reading for understanding. In *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, pages 1–6. IEEE.

Marcello Ferro, Claudia Marzi, Andrea Nadalini, Loukia Taxitari, Alessandro Lento, and Vito Pirrelli. 2024. Readlet: a dataset for oral, visual and tactile text reading data of early and mature readers. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13595–13609.

John Fox, Gregor Gorjanc Friendly, Spencer Graves, Richard Heiberger, Georges Monette, Henric Nilsson, Brian Ripley, Sanford Weisberg, Maintainer John Fox, and MASS Suggests. 2007. The car package. *R Foundation for Statistical computing*, 1109(1431):84039–5.

Jonatas Grosman. 2022. Fine-tuned XLS-R 1B model for speech recognition in Italian. https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-italian.

Albrecht W Inhoff, Matthew Solomon, Ralph Radach, and Bradley A Seymour. 2011. Temporal dynamics of the eye–voice span and eye movement control during oral reading. *Journal of Cognitive Psychology*, 23(5):543–558.

Diman Karagiozov, Svetla Koeva, Maciej Ogrodniczuk, and Cristina Vertan. 2011. Atlas–a robust multilingual platform for the web. In *Proceedings of the German Society for Computational Linguistics and Language Technology Conference (GSCL 2011)*, pages 223–226.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: Theory and practice in corpus design. *Journal of Language Modelling*, 0(1):65–110.

Jochen Laubrock and Reinhold Kliegl. 2015. The eye-voice span during reading aloud. *Frontiers in psychology*, 6:1432.

Alessandro Lento, Andrea Nadalini, Marcello Ferro, Claudia Marzi, Vito Pirrelli, Tsvetana Dimitrova, Hristina Kukova, Valentina Stefanova, Maria Todorova, and Svetla Koeva. 2024. Assessing reading literacy of Bulgarian pupils with finger–tracking. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)*, pages 140–149.

Claudia Marzi, Marcello Ferro, Andrea Nadalini, Vito Pirrelli, Maria Todorova, Tsvetana Dimitrova, Valentina Stefanova, Hristina Kukova, and Svetla Koeva. 2026. Comparable Reading Development in Bulgarian and Italian: Cross-Linguistic Insights from a Finger-Tracking Study. *Languages*, 11(4):70.

Claudia Marzi, Andrea Nadalini, Alessandro Lento, Manu Srivastava, Alice Todesco, Vito Pirrelli, and Marcello Ferro. 2025. Oral text reading as a multisensory task. *Lingue e linguaggio*, 24(1):141–156.

Claudia Marzi, Anna Rodella, Andrea Nadalini, Loukia Taxitari, and Vito Pirrelli. 2020. Does finger-tracking point to child reading strategies? In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 204–210.

Heidi Anne E Mesmer and Karen Lake. 2010. The role of syllable awareness and syllable-controlled text in the development of finger-point reading. *Reading Psychology*, 31(2):176–201.

Andrea Nadalini, Claudia Marzi, Marcello Ferro, Alessandra Cinini, Paola Cutugno, and Davide Chiarella. 2025. Inferential reading skills in high school: A study on comprehension profiles. *Education Sciences*, 15(6):654.

Andrea Nadalini, Claudia Marzi, Marcello Ferro, Loukia Taxitari, Alessandro Lento, Davide Crepaldi, and Vito Pirrelli. 2023. Eye-voice and finger-voice spans in adults' oral reading of connected texts: Implications for reading research and assessment. *The Mental Lexicon*, 18(3):366–400.

Vito Pirrelli and Svetla Koeva. 2024. Developing materials for assessing reading literacy and comprehension of early graders in Bulgaria and Italy. *Foreign Language Teaching*, 51(1).

Manu Srivastava, Marcello Ferro, Vito Pirrelli, and Gianpaolo Coro. 2026. Enhancing token boundary detection in disfluent speech. *Intelligent Systems with Applications*, 29:200614.

Loukia Taxitari, Claudia Cappa, Marcello Ferro, Claudia Marzi, Andrea Nadalini, and Vito Pirrelli. 2021. Using mobile technology for reading assessment. In *2020 6th IEEE congress on information science and technology (CiSt)*, pages 302–307. IEEE.

Hui Zhang. 2020. *The Distribution of Disfluencies in Spontaneous Speech: Empirical Observations and Theoretical Implications*. Doctoral dissertation, University of Pennsylvania.

Laoura Ziaka and Athanassios Protopapas. 2025. Dynamic adjustment of the eye–voice span and articulation duration in the course of multi-item naming tasks: Evidence for lockout scheduling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Appendix A Acoustic Annotation of Transcribed Data

A.1 Annotation labels

All non-word audio segments as well as pauses and annotators‘ comments were marked in triangular brackets <>.

Ex.

избязло <pronunciation mistake – избягало ’run’>

ро- <partial word>

в-няко-<unclear pronunciation>

A.2 Pause annotation

Annotations were grouped into functionally similar categories (e.g., restarts vs. repetitions). Minimum duration thresholds were also defined for coding speech pauses.

These annotations enable additional analyses of reading behaviour. Pauses were distinguished according to their position: Pauses were distinguished according to their position:

- constituent boundary (marking the end of a sentence, a clause or a phrase);
- internal constituent pauses (hesitations, repetitions, sighing).

A.3 Examples of pause annotation

Constituent boundary pauses:

Ex.

Тара.PROPN <pause1>, както се.REFL.CL <pause2> казваше.V <pause3> (3rd-grader)

Tara <pause1>, as (she was) REFL.CL <pause2> called, <pause3>

In the example, pause1 and pause3 mark a constituent boundary (of an embedded clause), while pause2 unexpectedly separates the reflexive clitic (се) from the verb (казваше ’named’), thereby resulting in the clitic being associated with the preceding wh-constituent.

Internal constituent pauses:

Ex.

благородна- <pause1> -та кралица на <pause2> паяците <pause3> в моята <pause4> градина <highly muted> <pause5> (3rd-grader)

noble- <pause1> -the queen of <pause2> spiders-the <pause3> in my-the <pause4> garden <highly muted> <pause5>

In this example, pause1 occurs before the definite article (-та), which may be analysed as a clitic-like element. Pause2 separates the preposition на (’of’) from its governed noun паяците (’the spiders’), while pause4 separates two noun phrases. Finally, pause3 and pause5 mark constituent boundaries.

A.4 Annotation observations

Observations made during the annotation may provide an additional perspective on the assessment of reading skills. They allow for the monitoring of a range of interrelated parameters, including reading speed and fluency, as well as comprehension of words in context and lexical breadth.